实体名称规范的研究探索[1]

刘建华[1,2]，郭红梅[1,2]
[1]（中国科学院文献情报中心 北京 100190）
[2]（中国科学院大学 北京 100190）

[摘要]本文以文本处理中的基本任务之一--实体名称规范为主题，阐明了实体名称规范中两种类型的任务，一个实体多个名称的实体共指消解问题和一个名称指代不同实体的实体歧义问题，结合这两类任务，综合分析了当前的相关研究成果，重点介绍了当前解决实体名称规范时典型的思路与方法，推动实体名称规范研究的重要的项目与重要评测会议，并结合当前研究中仍存在的问题，分析探讨了实体名称规范的研究趋势。

[关键词] 实体名称规范;实体消歧;大规模知识库;社会网络

Study on Named Entity Normalization

Liu Jianhua[1, 2]，Guo Hongmei1[, 2]

[1](National Science Library, Chinese Academy of Sciences, Beijing 100190, China)

[2](University of Chinese Academy of Sciences,Beijing 100190, China)

[Abstract]This article focus on the Named Entity Normalization (NEN),which is a basic task of text processing. It describes two types sub-tasks of NEN,Co-reference Resolution and Entity Disambiguation. Combined with the mentioned two su-tasks, it reviews current related research, introduces the typical methods, importance projects and evaluation conference closed to the theme. Besides, it analyzes the research trend of NEN based on current problems.

[Keywords]Named Entity Normalization; Entity Disambiguation; large-scale knowledge; social network

1 概述

在现实世界中，不同的人经常会给予同一个事物不同的名称或描述。随着信息科技的不断发展，网络资源越来越多，这类事物的名称也越来越多样化，这为计算机的自动理解和计算带来了很大的挑战。为了支撑相应的文本处理任务，如机器翻译、信息检索、数据挖掘等，将这些名称、描述与其对应的事物对应起来，并从中选择一种规范的表达作为不同名称或描述之间的核心关联非常有必要，由此产生了实体名称规范这样一个概念。

从主题角度而言，与实体名称规范密切相关的研究主题包括实体名称共指消解、缩略语识别、实体名称消歧等，其相应的英文名称为“Named Entity Disambiguation，Abbreviation Reorganization，Co-reference Resolution，Named Entity Normalization”等。从任务角度而言，实体名称规范包括两种类型的任务，（1）一个实体有多种名称的实体共指问题。该问题既包括代词的共指消解，如“he，she”等人称代词实际指称对象的查找，也包括名词性称呼的消解，如“44th Present of US”、“Barack Obama”、“Present Obama”等可能均指代同一个人，这就需要明确这些实体名称是否确定指代同一个实体概念。（2）一个名称可能指代不同的实体的实体歧义问题[1]。实体由于一个词义的表达方法（从含义的有限集合枚举到

[1]本文系中国科学院文献情报中心青年人才项目“基于开放 KOS 的领域主题学术关系网络扩展方法研究”项目（课题号：青 1303）的研究成果之一青 1303。

基于规则的新含义的产生）、含义列表的细粒度（从细微的区别到反义词）、面向领域的与非严格定义的自然文本等原因，往往会出现一个实体名称可以对应到多个命名实体概念上的问题，比如，“Washington”既可能指称华盛顿州，也可能指代美国的第一任总统。针对此，就需要明确这些实体名称具体是什么概念。

本文以实体名称规范为主题，综合分析、研究了当前的相关研究成果，重点介绍了当前解决实体名称规范时典型的思路与方法，推动实体名称规范研究的重要的项目与重要评测会议，并结合当前研究中仍存在的问题，分析探讨了实体名称规范的研究趋势。

2 实体名称规范的主要思路与方法

实体名称规范事实上主要是一个以计算的方式自动辨析词语在上下文中的真实含义的过程[2]，它与常规的词义消歧任务（Word sense disambiguation）有很多相似之处，但由于命名实体概念列表的缺乏、实体名称指称形式更为多样（全称、缩略语、别称、代词、简称、不同语系的拼写差异-英美语系等）等问题的存在，实体名称规范的任务更加复杂。要完成这样的任务，其中需要涉及到很多知识，不仅仅需要语言学方面的常用知识，如浅层的词汇、语法、句法等的分析，还需要用到很多语义及其背景知识信息。本文对当前的一些主要研究进行了梳理，提炼出三种主流的方法思路，具体阐述如下。

2.1 基于 Web 对象属性信息的实体名称规范研究

Web 页面中往往嵌入了各种各样的对象，如人、产品、组织机构等实体名称。从 Web 页中抽取并集成这些对象，可以实现功能强大的对象层内容揭示。此类方法的优势在于其来源数据的特殊性，这些来源于 web 网页的资源在获取其属性方面具有很大的便利性，从而为基于属性模板的共指消解提供了很大的便利条件。

Zaiqing Nie[3]等认为，Web 对象是描述某一 Web 信息的数据单元，通常可以看作是与应用领域相关的概念。一个 Web 对象可以通过一系列的属性表示，如 A={a1, a2,……, am}。对象的属性集可根据领域的需要预先设置。在实际研究中，Zaiqing Nie[4]等将 Web 上一系列有一定结构的相同条目（如产品列表、服务列表等）称为数据记录，先从数据源中抽取出与领域相关的数据记录，形成对象记录级别的标识。然后进行对象属性级别（attribute-level）的抽取，这一过程主要是上一步抽取出的数据记录进行分析，将数据记录中的不同部分标识成为不同的属性，并且从多个来源的记录中，实现同一对象不同属性值的获取。最终依据所获取的属性值来实现对象的融合。

尽管该类方法实现的便利性和准确习惯都较高，但该类方法也有较大限制，对于来源数据的格式限制较多，仅适用于少量结构化或半结构化描述了实体的网页。

2.2 基于大规模知识库的实体名称规范研究

实体消歧的关键问题是测度实体名称出现的相似度，传统的测度方法是利用 BOW（bag-of-word）模型，但它忽略了语义关系。随着网络上结构化、半结构化知识库的出现，为弥补以往方法的不足，不少学者提出了利用如 Wikipedia[5]、Yago[6]等资源库构建大规模的知识库，基于这些知识库提供的背景知识来提升实体名称规范的效果，这也是当前实体名称规范研究中的核心内容之一。

Wikipedia 由于覆盖概念多，每篇文章中都包含了一个实体或一个概念的信息，具有丰富的语义信息且内容时时更新等特点，往往成为研究者们在开展此类研究或构建其它大规模知识库时的首选。Anthony Fader[7]等介绍了 GROUNDER 系统，通过利用 Wikipedia 上用户贡献的信息和新的消歧模型，有效利用先验信息，组合先验信息和语境信息以提高消歧精度。Hien[8] T 等人将文本中提到的实体映射到 Wikipedia 中正确的实体，在基于候选实体统计秩序模型基础上，证明 Wikipedia 和文本的功能组合是消歧的最好选择。Danuta Ploch[9]等人将实体名称消歧看做是将文本中的实体提及与预定义在知识库中的指称词相关联的任务，他们在研究中通过挖掘共现的实体间在 Wikipedia 里的关联关系，通过实体共现与歧义形式的关系推导出可用于分类候选实体的功能范围，并将消歧功能进行组合，利用 SVM 分类器得到了有效的结果。

但是由于 Wikipedia 本身在数据的准确性、概念结构的表达方面仍存在不足，因此，不少研究者又将眼光转向了近年来的热门知识库之一 linked open data（LOD），经过人工筛选、组织过的 LOD 在准确性和关联表达方面具备更强的知识处理优势。Danica Damljanovic[10]等人认为 Linked Data 是扩充已可用语境的有效资源，并将先进的命名实体工具与基于 Linked Data 相似度测度方法进行结合，证明该方法能提高 Wikipedia 消歧精度。Kamel Nebhi[10]等人采用 FreeBase 和句法分析结合的方式完成词义消歧的任务，试验显示了消歧效果的提升。

除 LOD 外，各种语义层级关联更为丰富的本体也是研究者们探索实体名称规范的重要知识库。Horacio Saggion[11]等基于欧盟的 MUSING[2]平台，在跨数据源的知识单元获取与集成任务方面做出了一定探索，他们是整个研究过程分为两个部分，一是基于本体的信息抽取，二是基于本体的跨数据源对象集成。其中，由领域专家构建的商业本体是系统的首要特征，该 Ontology 包含商业领域的类层次结构、关系和属性；其定义的对象主要包括：公司名、公司雇员数目、公司地址、网址、电话、传真和盈利状况等。在对每一片文档进行标注后，获取各标注对象所在的文档和描述内容部分，计算其相似度，实现多数据源中同一个标识对象的聚类，从而实现命名实体的规范。Farhad Abedini[12]等人利用 YAGO 中提供的大量的实体之间的事实描述来鉴别文本中的语义实体。Xianpei Han[13]等人综合利用 WordNet、Wikipedia、网页信息等多种知识源挖掘实体指称项的上下文语义信息，并提出了基于图的知识表示模型，将异构语义信息融合在统一的基于图的知识表示框架下，以此为基础挖掘概念之间的潜在语义关联，从而同时集成来自于不同知识源的语义知识，有效提升了实体名称规范的效率。

2.3 基于社会网络的实体名称规范研究

随着搜索引擎和社会网络挖掘技术的不断发展，利用人物社会关系关联构建社会网络，进而实现相应的实体消解方法也逐渐成为目前的关键思路之一，此类方法通常主要应用于人名的消歧，通常是先使用谱聚类对社会网络中的人名聚类,然后根据不同社会网络边权值和不同图划分准则对人名消歧效果的影响，引入模块度阈值作为社会网络划分的停止条件[14]。

[2] MUlti-industry, Semantic-based next generation business INtelliGence，基于语义的下一代多产业商业情报

图 1 展示了典型的基于社会网路的实体名称规范框架。

在基于社会网络的实体名称规范方面，RonBekkerman[15]等人提出了一种非监督的框架来解决检索某个特定人物时返回大量无关人员页面的问题。其中两个关键内容包括网页间的链接关系与 Agglomerative 重复聚类。在该方法中，网页间的链接关系即主要用于构建人物的社会网络。郎君[16]等人依据同名的不同人物具有不同的社会网络的思想，利用检索结果中共现的人名发现并拓展检索人物相关的潜在社会网络，结合图的谱分割算法和模块度指标进行社会网络的自动聚类，在此基础上实现人名检索结果的重名消解. 在人工标注的中文人名语料上进行实验，整体性能达到较好水平，图聚类算法能帮助连通社会网络的进一步划分，从而提高消解效果。

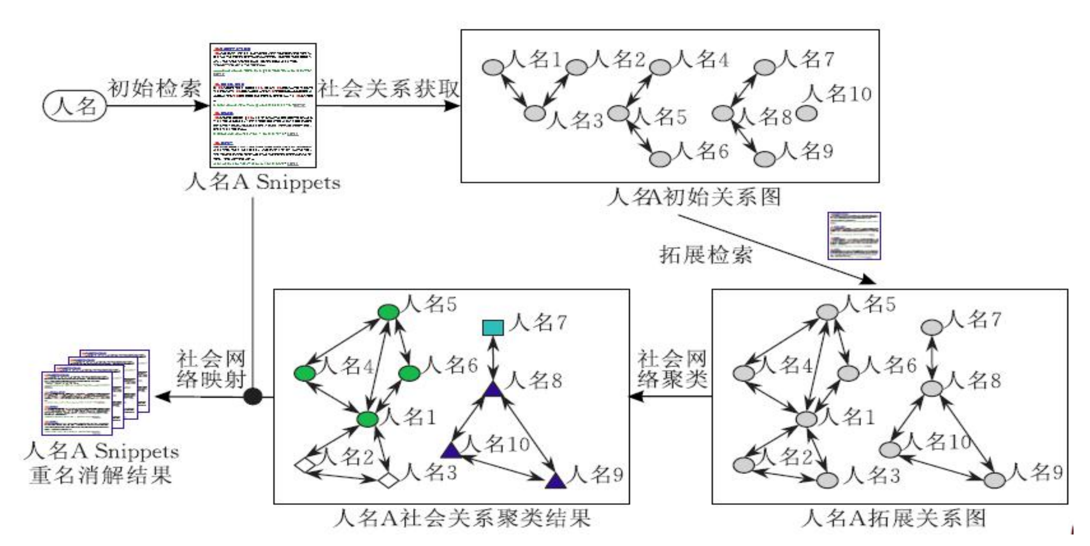


图 1 典型的基于社会网路的共指消解框架[16]

陈晨[14]等人先使用谱聚类对社会网络中的人名聚类,然后根据不同社会网络边权值和不同图划分准则对人名消歧效果的影响,引入了模块度阈值作为社会网络划分的停止条件，在共指消解方面取得了较好的效果。JADERICK P. PABICO[17]针对社交网络中的实体名称歧义问题，提出采用图-字图的方式来确定不同实体的相似性，从而解决实体名称的歧义。

Mohammad 等人针对数字图书馆中多来源数据汇集造成的作者名称消歧，提出通过构建共著网络，利用启发式聚类方法，

3 实体名称规范相关的重要项目、评测会议

实体名称规范的研究离不开重大项目、国际评测会议的推动发展，本文对这些重点内容进行了梳理，希望给后续的研究者提供相应的参考。

3.1 国内外主要的实体名称规范项目

（1）英国国家档案馆 TNA-Search 项目[18]

英国国家档案馆 TNA[3]（the National Archives）是大规模实体名称规范的代表性项目。

[3] TNA-search 作为 Govenment Web Archive Project 中的一部分，主旨在于如何用简单直观的机制，提高 TNA 中与政府网站相关的记录（记录回溯到 1997 年，包含了大概 7 亿的网页）的开放利用度

为了解决项目中的实体名称规范问题，TNA-Search 项目主要利用 GATE，联合了 FactForge[4] 和 SKB（Semantic Knowledge Base）Ontology[5]，构建了大规模的语义仓储库（Large knowledge base，LKB），通过仓储库所提供的详细的对象描述等背景信息，计算实现实体名称的规范。

具体而言，该项目基于 LKB 直接将文档中的实体与各种不同的本体建立关联，或者通过其中的实例，或者通过概念。LKB 使用了一系列 SPARQL 查询集合的配置文件到 SKB 中检索。标注的实体与 SKB 中的实例关联是通过两个互补的途径完成的：通过 LKB 词典找到一个匹配时，SKB 中类与实例信息被添加到文本中的相关实体上；文本中的实体与 SKB 中的类或实体没有直接关联时，通过共指的方式实现关联。即如果文本中某段提及在上述过程中已经与 SKB 建立关联时，该实体所有共指提及均可通过 TNA Instance Generator 自动获得相同类和实例信息。在进行规范标注时，项目将一篇文档中同一个实体的不同表达关联在一起，同时还添加通过 semantic tagger 发现的标注间的特征关系。通过这种规范标注方式，TNA-Search 实现了人物、地理名称、机构、时间等 11 种命名实体的自动标注与规范。

（2）OKKAM[19]

OKKAM 是由欧盟委员会资助的第七框架项目(FP7)下的一个大规模集成项目,其基本理念是根据 14 世纪的“奥卡姆剃刀(Occam’s razor)”原则，提倡如果没有必要则不增加实体的标识符。OKKAM 为内容创建者、编辑和开发人员等提供一个全球性的基础设施,称为实体命名系统(entity name system，简称 ENS)，该系统中包含了一种基于特征的实例匹配方法 FBEM，该匹配方法通过集成两个实例标识符的多种不同特征属性及其属性值之间的相似度，识别出可能的对象共指。例如，FBEM 使用了基于 Levenstein 编辑距离的方法来比较实例标识符的本地名。

（3）国内典型的项目

共指消解和实体消歧是文本处理中非常重要的任务之一，它对于提高信息检索的效率、深度的文本挖掘有着非常重要的作用，国内目前在此方面也有不少相关的研究项目在开展。比较典型的有清华大学的 RiMoM[20]和南京大学的 ObjectCoref[21]。

RiMOM 是清华大学研发的一种集成了多种本体匹配方法的多策略本体匹配系统，其中也包含了多种实例匹配方法。针对实例匹配，RiMOM 将每个实例所含信息分为 6 类：URL、元信息、名称、字符串类型信息、非字符串类型信息和邻居信息。通过基于编辑距离的方法和向量空间模型，计算实例所含各种信息之间的相似度，并使用元信息和非字符串类型信息进一步过滤,最后通过多种策略将各种相似度集成起来用于发现对象共指。

与 RiMoM 不同，南京大学的 ObjectCoref 基于语义 Web 搜索系统 Falcons 提供的数据集，目前已经包含 7300 多万个实例标识符。ObjectCoref 首先利用语义等价推理，包括 owl:sameAs、函数型或反函数型属性以及基数或最大基数限制，构建出一个初始训练集；随后，基于这个训练集不断学习,自举式地识别对象共指，其中的关键技术是从训练集中学习

---

[4] OntoText 开发的知识库，该知识库包含了超过 22 亿声明和来源于多个源的数据集

[5] OntoText 开发，基于 CGO（Central Government Ontology）与 UK 政府的官员职位、8138 个官员名字以及无歧义的 UK 政府机构名称

出最适合识别对象共指关系的属性及属性值。该系统还考虑了频繁属性组合，同时使用两个属性识别对象共指(例如经度和纬度、 姓和名),进一步提高消解的准确度。另外，还基于语义等价关系是否可以解引以及实例标识符在不同 RDF 文档中的出现次数等，对共同指称同一对象的实例标识符进行排序。ObjectCoref 提出了一种新的语义等价推理与相似度计算相集成的体系结构，能够较为全面地识别对象共指，但是训练集中的错误共指关系可能会导致学习过程中的错误积累。使得识别的准确性降低。

3.2 实体名称规范的相关评测会议

为了促进实体名称规范研究的不断发展，国际上有不少与之相关的评测会议，这些评测会议通过细化评测任务，提供相应的语料集合，提供交流的平台，推动者相关研究的不断发展。本文筛选了几个比较典型的评测会议进行了介绍，以期为其它研究提供一些参考。

（1）Automatic Context Extraction(ACE)与 Text Analysis Conference（TAC）

ACE 会议是从 1999 年 7 月开始酝酿，2000 年 12 月正式启动，由美国国家安全局（NSA），美国国家标准和技术学会（NIST），以及中央情报局（CIA）共同主管，到今年为止已经举办过 8 届[22]。ACE 的测评任务定义为：实体探测与识别(Entity Detection and Recognition ，简称 EDR)、价值探测与识别(Value Detection and Recognition ，简称 VAL)、时间表达识别与标准化(Time Expression Recognition and Normalization ，简称 TERN)、关系探测与识别(Relation Detection and Recognition ，简称 RDR)以及事件探测与识别(Event Detection and Recognition，简称 VDR)。共指消解的评测任务主要蕴含于实体探测与识别 EDR 中。该任务将篇章中出现的各种提及表述指向对应的实体，从而给出一个实体全面的描述。这项任务中首先需要识别出各种表述，然后将描述同一实体的表述合并，该合并过程就是共指消解的过程。值得一提的是，从 2003 年开始 ACE 中开始包含中文的相关评测，至今已经开展 5 次评测。其中的共指消解也是迄今为止唯一的中文共指消解国际评测。

在 2008 年之后，ACE 会议被 Text Analysis Conference（TAC）[23]会议所取代，TAC-KBP 从 2009 年开始到现在共进行了六届，该评测任务中直接与实体名称规范相关的即实体链接(Entity Linking)评测。目前，TAC 实体链接任务的目标实体知识库使用 2008 年 10 月版本的 wikipedia 构建，包含近 82 个实体，其中有人物实体 11 万，组织实体 5.5 万，地理实体 11 万，其它类别实体 53 万，目标知识库总量约 2.6G[24]。

（2）web 环境中人名消歧任务评测会议-Web People Search Evaluation（WePS）

WePS 是针对英文网页中人名消歧任务进行评测的一个专门会议，由 Julio Gonzalo 和 Satoshi Sekine 主要负责组织，至今为止共组织过三次[25]。该任务集中于在 web 检索场景中人名的消歧。参加测试的系统将在接收到一个以人名为检索式的 web 检索后，确定有多少个不同的涉及的人员在检索结果中，并将特定的指称分配给相应的文档。从总体上来说，这个任务是个聚类问题。对给定的一组文档，按照文档中出现的某个指定的人名所指向的人进行聚类。最后，在每个类中，所有指定的人名都必须是指向现实生活中的同一个人。从 WePS3 发布的评测任务看，在该评测中，需要重点从人物的属性角度出发，包括人员的生日、出生地、别名、工作、所属机构、获得奖项、学校、学位、专业、民族、电话等多个方面年代信

息。受该项目启发，李文捷等人也于 2010 年组织发起了专门针对中文人名消歧的评测任务[26]，至今已经举办了两届。

（3）指代消解练习（ARE）[27]

2006 年 11 月到 2007 年 3 月，英国伍尔佛汉普敦大学发起了一个名为指代消解练习 (ARE)的共指消解评测。这项评测是在英文上进行的迄今为止最全面的共指消解评测，包含四项评测任务:

- 预标注文档上的人称代词消解：文档内的名词短语都被识别出来，而且需要消解的代词也被标注出来。参加系统需要对每个人称代词在一个不包含人称代词的名词短语列表中找到正确的先行语。
- 预标注文档上的共指消解：文档内所有的名词短语都被识别出来，参加系统需要将文档内的所有共指链识别出来。
- 生语料上的人称代词消解：和第一项任务不同的是，评测文档没有经过任何标注，需要参加系统自行识别相关信息。
- 生语料上的的共指消解：和第二项任务不同的是，评测文档没有经过任何标注，
- 需要参加系统自行识别相关信息。

除上述的三种不限于领域的评测外，还有一些领域特定的共指消解任务评测，如生物医药领域的生物医药领域的自然语言处理及应用联合工作组 JNLPBA(Joint Workshop on Natural Language Processing in Biomedicine and Its App lications) 和以及生物学领域信息抽取的关键评价 BioCreAtIve(Citical Assessment of Information Extraction Systems in Biology)。这些评测会议不断推动着实体名称规范研究的开展。

4 实体名称规范的研究趋势

尽管目前针对实体名称规范的研究已经开展得较为成熟，但从评测会议的结果（2012 年参加 TAC entity linking 测评的系统平均效率为 72.1%[28]）来看，目前的识别效率仍不足以满足大规模的实际应用，其中还面临着很多问题需要解决。比如空目标实体问题、知识库的覆盖度问题、知识库不确切的问题、知识库使用的问题等等[29]。因此，围绕着这些问题，此领域的研究主要存在以下几种发展趋势。

（1）算法趋于多模型的融合

在过去的研究中，基于语言学特征的统计学方法和机器学习方法主流是分开思考的，很多研究都是在机器学习的分类或聚类中选择特征是再考虑加入一些语言学特征，这种融合方式对提高识别的效率比较有限。目前的研究中，研究者们逐渐开始考虑利用语言学思路来构建更加丰富的机器学习模型。Elango 提出了一种初始化的建议，结合中心理论和条件随机域模型(CRF)来实现人称代词消解。基于 CRF 模型的灵活性，依赖于上下文的传递优选性能被很好的融入到模型中[30]。Poesio[31]等人将子句作为话语单元，将篇章可以表示成一系列子句的集合，进而将篇章表示为一系列预指中心集合的特征空间。这个预指中心列表构成的特征空间可以融合一些相关特征，例如语法角色、性别、单复数等。类似的序列 CRF 模型上的推理和估计，还可以采用 Sutton and McCallum 讨论的技术[32]。

（2）消歧特征的筛选越来越多样化

从当前发表的研究论文集中的研究主题上看，研究者们越来越重视在实体名称规范中引入越来越多的特征，单纯从算法上进行改进而实施基于“知识匮乏”的研究方法越来越不被主流研究所看重。归纳起来，目前常用的实体消歧特征主要如表 1 所示。

表 1 实体消歧特征归纳

特征大项	具体消歧特征
词汇特征	单复数、距离、人称、字符串匹配、词性等
语法特征	句法依存、语法角色等
语义特征	实体类型、实体属性（不同类型实体的属性定义各有不同）、同位、别称、维基百科类别的重叠度等
其它	实体句内共现、上下文相似度、维基百科文章中的入链及出链文本等

被应用的特征越来越多，而不断涌现出的各种语料资源库恰恰为这些深层的语言学知识获取提供了非常好的途径。这些知识主要可以从以下三种途径获取：①常规的知识库，如WordNet、HowNet、WikiPedia、DBPedia、Yago等；②利用大规模的语料库挖掘模式信息。如Hearst等通过构建了“is-a”等模板，用于从文本中发现同义词[33]；Bergsma[34]在一个经过Minipar依存分析的语料库上获取了大量的指代信息，实现了英文名词短语性别和单复数信息的模板化提取；Yang and Su[35]利用语料库中发现的模板信息来增强共指消解。③充分利用互联网这一语料库，利用搜索引擎显示的各个产寻得到的返回数来计算各种相关信息。第三种方法是将整个互联网当成一个巨大的语料库，利用搜索引擎显示的各个查询得到的返回数来计算各种相关信息，例如Poesio等人通过计算互信息来考察两个短语的关联程度。

（3）大规模知识库的自动构建成为实体规范研究的重要组成之一

实验充分表明，高质量的大规模知识库对提升实体名称规范的效率有很强的支撑作用。面对当前指数级增长的网络数据，依靠人工的专家构建知识库方式显然费时费力，且会造成信息的滞后。因此，富含语义信息关联的大规模知识库的自动构建显得尤为重要。开放式信息抽取技术的研究以及 wikipedia、freebase 等大规模半结构化的网络知识库的出现，为大规模知识库的自动构建提供了良好的基础。目前，较有代表性的工作有基于 Wikipedia 的 YAGO，该语料库采用实例、实例间关联三元组的方式存储知识，所有的实例和实例间的关系均来源于 wikipedia 的 category pages，并与 WordNet 进行衔接，对于每一个实体事实 YAGO 还赋予了可信度的标注，准确率达到 95%。YAGO2 中包含了 1000 万个实体及 1.2 亿条描述实体关联的事实记录[6]。此外，中国科学院自动化所的赵军等人，利用在信息抽取方面的技术积累，以《中国大百科全书》知识体系作为目标知识库的结构，从网络知识库中抽取概念实例并综合利用网络百科网页中蕴含的丰富的语义标签、半结构化信息和非结构化信息进行概念实例挂载，将百科知识库从 8 万条目扩展为百万条目级别，在此基础上进行概念属性抽取，为下一步研发面向开放式的自动问答系统提供了知识资源的支撑。

5 结语

本文围绕国内外与实体名称规范相关的理论、方法进行了深入广泛深入的分析，分别从实体名称规范的主要思路与方法，目前国内外典型的几个实体名称规范项目和评测会议，深入了解实体名称规范的主要内容，并结合实体规范研究面临的现实问题，分析了实体名称规

范的研究趋势。

参考文献

[1] Hien T. Nguyen1, Tru H. Cao. A Knowledge-Based Approach to Named Entity Disambiguation in News Articles[J]. AI 2007, LNAI 4830, pp. 619–624, 2007
[2] ROBERTO NAVIGLI. Word Sense Disambiguation：A Survey[J].ACM Computing Surveys, Vol. 41, No. 2: 10-69
[3] Nie, Z., et al., Web object retrieval[C].In: Proceedings of the 16th international conference on World Wide Web, 2007: 81-90
[4] Nie, Z., et al., Object-level ranking: bringing order to Web objects[C].In: Proceedings of the 14th international conference on World Wide Web, 2005: 567-574
[5] Wikipedia.http://www.wikipedia.org/[EB/OL] (Accepted:2014-11-26)
[6] YAGO2s: A High-Quality Knowledge Base. http://www.mpi-inf.mpg.de/departments/databases-and-information-systems/research/yago-naga/yago/[EB/OL] (Accepted:2014-11-26)
[7] Fader, A., Soderland, S., Etzioni, O.: Scaling Wikipedia-based named entity disambiguation to arbitrary web text[C]. In: Proceedings of the IJCAI Workshop on User-contributed Knowledge and Artificial Intelligence: An Evolving Synergy, Pasadena, CA, USA, pp. 21-26 (2009)
[8] Hien T. Nguyen, Tru H. Cao Exploring Wikipedia and Text Features for Named Entity Disambiguation[J].Intelligent Information and Database Systems Lecture Notes in Computer Science, 2010, Volume 5991/2010: 11-20
[9] Danuta Ploch. Exploring Entity Relations for Named Entity Disambiguation[C]. In: Proceedings of the ACL 2011 Student Session,Portland, OR, USA,2011
[10]Danica Damljanovic, Kalina Bontcheva.Named Entity Disambiguation using Linked Data[EB/OL]. http://2012.eswc-conferences.org/sites/default/files/eswc2012_submission_334.pdf(Accepted:2014-11-26)
[11] Horacio Saggion,Adam Funk,Diana Maynard,Kalina Bontcheva. Ontology-based Information Extraction for Business Intelligence[EB/OL].https://gate.ac.uk/sale/iswc07/musing/musing-iswc07.pdf(Accepted:2014-11-26)
[12] Kamel Nebhi. 2013. Named entity disambiguation using freebase and syntactic parsing[C]. In:Proceedings of the First International Workshop on Linked Data for Information Extraction (LD4IE 2013) co-located with the 12th International Semantic Web Conference (ISWC 2013)
[13] Xianpei Han. Named Entity Disambiguation by Leveraging Wikipedia Semantic Knowledge[C]. In:Proceedings of the 18th ACM conference on Information and knowledge management,2009:215-224
[14] 陈晨，王厚峰.基于社会网络的跨文本同名消歧[J].中文信息学报 ，2011(5)：75-82
[15] Ron Bekkerman, McCallum Andrew. Disambiguating web appearance of people in a social network[C]. In WWW '05 Proceedings of the 14th international conference on World Wide Web，2005:463-470
[16] 郎君，秦兵等.基于社会网络的人名检索结果重名消解[J].计算机学报,2009(7)：1-10
[17]JADERICK P. PABICO. An Analysis of Named Entity Disambiguation in Social Networks[J]. Asia Pacific Journal of Multidisciplinary Research,2014(2):31-38
[18]Diana Maynard, Mark A. Greenwood. Large Scale Semantic Annotation,Indexing and Search at the National Archives[EB/OL]. https://gate.ac.uk/sale/lrec2012/tna/tna.pdf(Accepted:2014-11-26)
[19] Paolo Bouquet，Themis Palpanas，Heiko Stoermer，Massimiliano Vignolo. A Conceptual Model for a Web-scale Entity Name System[EB/OL]. http://www.inf.unibz.it/krdb/events/swap2010/paper-19.pdf (Accepted:2014-11-26)

[20] Li JZ, Tang J, Li Y, Luo Q. RiMOM: A dynamic multistrategy ontology alignment framework. IEEE Trans. on Knowledge and Data Engineering, 2009,21(8):1218−1232
[21] ObjectCoref.http://ws.nju.edu.cn/objectcoref/[EB/OL]. (Accepted:2014-11-26)
[22] Automatic Content Extraction (ACE) Evaluation.http://www.itl.nist.gov/iad/mig/tests/ace/[EB/OL]. (Accepted:2014-11-26)
[23] Text Analysis Conference.http://www.nist.gov/tac/[EB/OL]. (Accepted:2014-11-26)
[24]KBP 2013 Entity Linking Task Description V1.0. http://www.nist.gov/tac/2013/KBP/EntityLinking/guidelines/KBP2013_EntityLinkingTaskDescription_1.0.pdf[EB/OL]. (Accepted:2014-11-26)
[25] Javier Artiles， Andrew Borthwick， Julio Gonzalo， Satoshi Sekine， nrique Amigo. WePS-3 Evaluation Campaign: Overview of the Web People Search Clusteringb and Attribute Extraction Tasks[EB/OL]. (Accepted:2014-11-26)
[26] CLP2012-Chinese Language Processing.http://www.cipsc.org.cn/clp2012/bakeoff-cn.html[EB/OL]. (Accepted:2014-11-26)
[27]Constantin Or˘asan, Dan Cristea, Ruslan Mitkov, Ant´onio Branco, Anaphora Resolution Exercise: An overview.http://www.lrec-conf.org/proceedings/lrec2008/pdf/713_paper.pdf[EB/OL]. (Accepted:2014-11-26)
[28] Jeffrey Dalton,Laura Dietz.A Neighborhood Relevance Model for Entity Linking. http://ciir.cs.umass.edu/~dietz/entitylinking/oair2013.pdf[EB/OL]. (Accepted:2014-11-26)
[29] 赵军，刘康，周光有，蔡黎.开放式文本信息抽取[J].中文信息学报，2011（6）：98-110
[30] P. Elango. Coreference resolution: A survey. Project report of the course "Advanced natural language processing"[D], In computer science departments, university of Wisconsin Madison,2006
[31] M. Poesio，M. Kabadjov.A general-purpose, off-the-shelf anaphora resolution module: Implementation and preliminary evaluation[C]. In: Proc. of the 4th International Conference on Language Resources and Evaluation. Lisbon, Portugal
[32]C. Sutton，A. McCallum. 2006. An introduction to conditional random fields for relational learning[C], In: L. Getoor and B. Taskar, eds. Introduction to statistical relational learning: MIT Press
[33] M.A. Hearst. Automatic acquisition of hyponyms from large text corpora.[C].In:Proceedings of the 14th International Conference on Computational Linguistics, 1992
[34] S. Bergsma. Automatic acquisition of gender information for anaphora resolution[C]. In: B. Kégl and G. Lapalme eds. Canadian Conference on AI,2005,Victoria, Canada: Springer-Verlag, 342-353
[35] X. Yang，J. Su. Coreference resolution using semantic relatedness information from automatically discovered patterns[C]. In: J. Carroll, A. Bosch, and A. Zaenen eds. Proc. of the 45th Annual Meeting of the Association of Computational Linguistics}. Prague, Czech Republic: Association for Computational Linguistics, 528-535

作者简介：
刘建华（1984-），女，江苏南通，中级，在读博士，主要从事文本挖掘、信息抽取研究。
郭红梅（1985-），女，河南周口，在读博士，主要从事文本挖掘，科学计量相关研究。